Predictive Visual Analytics for Machine Learning Model in House Price Prediction: A Case Study

Norhayati Yahya, Norziha Megat Mohd Zainuddin, Nilam Nur Amir Sjarif, Nurulhuda Firdaus Mohd Azmi

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Jalan Semarak, 54100 Kuala Lumpur, Malaysia
yat.yahya@yahoo.com, , norziha.kl@utm.my, nilamnur@utm.my, huda@utm.my

Article history

Received:
14 May 2021

Received in revised form:
20 May 2021

Accepted:
30 May 2021

Published online:
26 June 2021

*Corresponding author:
yat.yahya@yahoo.com

Abstract

As an individual, buying a house is a nerve-racking process. It requires a huge amount of money, time-consuming and relentless worry whether it is a good deal or not. The uncertainty in the housing market and the motivation to own a house have raised questions among homeowners and buyers regarding how accurate the house prices can be predicted, and what attributes or factors influenced the house prices. There were studies conducted in Malaysia that applied machine learning in predicting house prices. However, most of the studies using the Valuation and Property Service Department (VPSD) dataset were conducted in different states, namely Selangor, Kuala Lumpur, and Johor. Thus, there is an opportunity to extend the study to predict the house price in Penang state, Malaysia due to the increase in house prices in Penang is the highest among all the states in Malaysia. Therefore, this study aims to produce a machine learning predictive model using 2,666 terrace houses actual property transactions in Penang from VPSD from January 2018 until December 2019. The dataset is split into a train-test (estimation-validation) set with 80% train set and 20% test set (80:20) proportion and separated by two groups of different feature selection dataset which is all feature and selected features. Hence, to capture the different performances from both groups. The predictive model development using Multiple Linear Regression, Random Forest, and K-Nearest Neighbors algorithms with different parameters. The predictive model's performance was evaluated based on error measurement metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). Its reveals that Random Forest of 250 trees using all feature dataset has been chosen as the best model which produces 23,786.856 for Root Mean Square Error (RMSE), 13,769.965 for Mean Absolute Error (MAE), and 4.674% Mean Absolute Percentage Error (MAPE) from the train set.

Keywords: *Predictive visual analytics; machine learning; predictive model; house price prediction*

1. Introduction

Highlighted by [1], most of the individuals who live in Asia have a major goal and are eager to become a homeowner. This goal is driven by the attractive and positive investment return of the house, unlike other assets. For example, in Malaysia, the transaction volume from 1990 to 2007 of the residential property has contributed more than 60% of the country's growth, which makes this industry the leading contributor. Basically, for the homeowner or investor, there are two types of prospective returns from buying houses, namely rental payment and capital gain from the increasing value in the property [2]. Moreover, investing in housing commonly seemed to be smart and beneficial for the owner due to their equity value

does not reduce vigorously [3]. Thus, created a decent prospect to increase wealth for the investors or homeowner.

However, the uncertainty in the real estate market has raised questions among homeowners and buyers regarding how accurate the house prices can be predicted and what attributes or factors affecting house prices [4]. Furthermore, as an individual, buying a property is difficult because the unfamiliar legal term involved in the process leads to the difficulty for a buyer to purchase houses directly from the seller [5]. Thus, the buyer and seller rely on the real estate company or agent as the middleman between them for the asset transfer procedure as well as to get information on the current house prices [6]. Moreover, most house buyers turning to online research as a popular information source to estimate the house price. However, both the Internet and real estate agent or company listed the houses in the website with increased prices which are different from one to another regarded for similar houses [7]. It shows that the buyer does not have the information on the current house price whether it is overvalued or undervalued.

According to [8], compared to statistical methods such as Hedonic Price Method (HPM), Fuzzy Logic System (FLS), and Analytic Hierarchy Process (AHP), machine learning (ML) produces higher accuracy performance. Due to that, the presence of a house price predictive model based on the ML technique has been very appealing to many property valuation professionals. For an equal reason, scholars have started to applied a variety of ML techniques to predict house prices [8]. Table I reveals the number of publications from the last decade by conducting a quick search in Google Scholar with a keywords of 'machine learning', 'house price', and 'prediction' [9]. Table 1.1 shows that there are few publications from 2010 until 2015. However, from 2016 until March 2020, there is rapid growth. It is clear that 'machine learning', 'house price' and 'prediction' is an important topic.

This study aims to produce a Penang's terrace house price predictive model from ML using the Valuation and Property Service Department (VPSD) dataset. It can give an idea as well as advice for the buyer to negotiate the price, especially for first-time buyers with relatively little experience, and advice purchasing strategies for buying properties. Besides, this research compared three different ML approaches which are, information-based learning, similarity-based learning, and error-based learning, and observe the differences in terms of model performance based on Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).

Table 1.1 Publication related to housing price prediction using machine learning

Year	No. of Publication
2010	24
2011	35
2012	38
2013	37
2014	83
2015	81
2016	117

Year	No. of Publication
2017	201
2018	300
2019	486
March 2020	90

The remainder of the study is organized as follows; Section 2 provides the reviews of related works that have been done in predicting house prices using ML. Then, the research methodology of the study is presented in Section 3. This will be followed by model implementation and the experimental results in Section 4. Finally, the conclusion of the study is provided in Section 5.

2. Literature Review

The house price prediction has already been studied using ML. The ML technique is constantly able to precisely see the potential by considering the significant factors. Subsequently, the predictive accuracy of housing models has gained much attention among scholars and has been widely studied. But, to apply the ML techniques in the real estate industry, the determination of substantial factors is important. It requires a pre-processing task, examination, and understanding of the collected datasets. According to [8], the accuracy of the results generated by the ML model is very much reliant on the dataset pattern, the parameter tunings, and the feature selections.

Throughout the year, various ML studies have been conducted to predict housing prices. Table 2.1 shows some of the different ML techniques use in developing the predictive model in predicting house prices. From the table, the fourteen studies listed the implementation of a predictive model based on the ML technique. A variety of ML techniques has been employed to build a predictive model of house prices. Either by the employment of individual ML techniques or by combining (ensemble) several ML techniques to enhance the performances of the predictive model. The comparison between ML techniques to get the highest performance result of the predictive model also is widely studied. For example, four ML techniques have been applied to produce the predictive model to predict the house prices in Montreal, Canada such as Multiple Linear Regression (MLR), Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) [10]. Another example is by comparing the performance of the predictive model developed based on five ML techniques; MLR, Ridge Regression, Lasso (Least Absolute Selection and Shrinkage Operator) Regression, RF, and Decision Tree (DT) to predict the Petaling Jaya, Selangor house prices [8].

Table 2.1 The summarization and comparison of fourteen previous studies in house prediction using machine learning

Article, Author, and Year	Sample Data	Algorithm	Validation Method and Evaluation Measurement	Result	Strength/ Limitation
[3]	2,000 real estate housing data.	• Logistic Regression • Decision Tree • Support Vector Machine	• R-squared • Root Mean Square Error (RMSE) • Mean Absolute Error (MAE) • Mean Square Error (MSE) **Validation:** 80:20 train-test split ratio	• Less error value, high accuracy, and R-squared values captured by Decision Tree with topmost accuracy of 84.59%, R-squared value of 0.98, RMSE value of 217, MSE value of 47,184.93, and MAE value of 5.68.	• The use of Logistic Regression (a classifier) is unsuitable because it gives a categorical output. • The learning algorithm compares information-based learning and error-based learning. • A comprehensive comparison of the predictive model performances using five error metrics measurements. • The sample size used in this study is big.
[6]	20,000 real estate housing data. Collected from UCI machine learning repository.	• Support Vector Machine • Gradient Boosting Regression Tree • Artificial Neural Network • Bagging • Multiple Linear Regression • Random Forest	• R-squared • Root Mean Square Error (RMSE) **Validation:** RF and MLR – 75:25 train-test split ration SVM and Gradient Boosting Regression Tree - 10-fold cross-validation ANN and Bagging – 70:30 train-test split ratio	• The lowest error provides by Random Forest with an RMSE value of 0.012 and 90% accuracy.	• Different validation methods are used for each algorithm which leads to inconsistent in the evaluation results. • A learning comparison of predictive models is from the information-based learning, and error-based learning. • The performance evaluation using two measurement metrics of error. • The sample size used in this study is big.
[8]	Housing selling prices for the year 2016 in the Petaling Jaya area, Selangor.	• Random Forest • Decision Tree • Multiple Linear Regression • Lasso Regression • Ridge Regression	• R-squared • Root Mean Squared Error (RMSE) **Validation:** 80:20 train-test split ratio	• The buying price has the strongest correlation with the Selling price with a 0.73 coefficient correlation value. • For all 4 groups of features, Random Forest has better fitted the model followed by Decision Tree. • Confirmed that no extreme difference between all groups regardless of strong or weak groups. • Verify that 'Buying' prices can be used for predicting Selling prices without considering other features.	• The learning algorithm compares five algorithms which are from error-based learning. • A comparison of the predictive model performances using two error metrics measurements. • 19 variables or house attributes were used as independent variables to predict house prices.

Article, Author, and Year	Sample Data	Algorithm	Validation Method and Evaluation Measurement	Result	Strength/ Limitation
[11]	16,601 apartment transaction data from the period of 2006 to 2017 in the district of Gangnam, South Korea. Provided by the Ministry of Land, Infrastructure, and Transport (MOLIT), South Korea.	• Random Forest • Hedonic Price Model (HPM)	• Mean Absolute Percentage Error (MAPE) • Coefficient of dispersion (COD) • R-squared **Validation:** 90:10 train-test split ratio	• The MAPE value for Random Forest is 5.5% compared to HPM with 20%. • The probabilities that the RF prediction was within 3%, 5%, and 10% of the actual market price were 53.5%, 72%, and 90.3%, respectively. • The HPM prediction are 10.4%, 17.4%, and 34.6%, respectively. • The probability of the RF predictions deviating more than 50% from the actual price was found to be only 0.5%, while the HPM predictions value almost 3.8%.	• The sample undergoes 10 trials for each measurement, and the result is obtained by averaging the 10 trials to eliminate the possibility that the results occurred by chance. • A comparison of predictive models between two models; Hedonic Price Model and information-based learning algorithm. • The comprehensive performance evaluation using three measurement metrics of error. • The sample size used in this study is big.
[12]	50 housing prices in Tadepalligudem of West Godavari District in Andhra Pradesh of India.	• Decision Tree Regression • Multiple Linear Regression	• Root Mean Square Error (RMSE) • Mean Absolute Error (MAE) • Mean Square Error (MSE) **Validation:** 80:20 train-test split ratio	• The Multiple Linear Regressions has outperformed the model than Decision Tree Regression in predicting the prices of the houses with an MAE value of 1.953, MSE value of 6.065, and RMSE value of 2.463.	• The sample size used in this study is too small which can affect the model performance evaluation. • The learning algorithm compares information-based learning and error-based learning. • The performance evaluation using three measurement metrics of error.
[13]	34,857 Melbourne housing market data from the year 2016 to 2018. Obtained from the Kaggle website.	• Support Vector Machine • Polynomial Regression • Regression Tree • Artificial Neural Network • Linear Regression	• Mean Square Error (MSE) **Validation:** 10-fold cross-validation	• The combination of Stepwise and tuned SVM, which produces the lowest error on this dataset, is the most competitive model with an MSE value of 0.0561.	• Data reduction and transformation such as PCA and Stepwise are applied to the dataset to achieve an optimal solution. • The learning algorithm compares information-based learning and error-based learning. • The performance evaluation using one measurement metric of error. • The sample size used in this study is big.
[14]	2,325 double story sale transactions for the year 2000 to 2016 in Mukim Pulai, Johor Bahru. Collected from Valuation and Property Services Department Johor Bahru (VPSDJB	• Artificial Neural Network	• R-squared • Mean Absolute Deviation (MAD) • Mean Absolute Percentage Error (MAPE) • Root Mean Squared Error (RMSE) **Validation:** 90:10 train-test split ratio	• The ANN model for Taman Mutiara Rini and Taman Bukit Indah produced high R-squared with low values for MAD, RMSE and MAPE. • The results suggest that models with large sample sizes (Sets 1 of Taman Mutiara Rini and Taman Bukit Indah) have superior performance compared to models with small sample sizes (Sets 2 of Taman Mutiara Rini and Taman Bukit Indah).	• The study implemented one learning which is from error-based learning. • A comprehensive comparison of the predictive model performances using four error metrics measurements. • Limited sample size which is 193 records used for training and 22 records were used for test and validation.

Article, Author, and Year	Sample Data	Algorithm	Validation Method and Evaluation Measurement	Result	Strength/ Limitation
[15]	7,023 second-hand housing in Shanghai, China.	• Support Vector Machine • Artificial Neural Network • Random Forest • Ordinary Least Square	• Root Mean Square Error (RMSE) • Mean Absolute Error (MAE) • Mean Absolute Percentage Error (MAPE) **Validation:** 80:20 train-test split ratio	• Support Vector Machine has the best prediction effect and has strong stability with an RMSE value of 0.865, MAE is 0.380, and MAPE is 0.093.	• The categorical variables are converted into numerical for statistical purposes. • A comprehensive comparison of predictive models from the Artificial Intelligence framework and machine learning. • The comprehensive performance evaluation using three measurement metrics of error. • The sample size used in this study is big.
[16]	The real house prices in the Pedurungan Sub-district of Semarang City, Indonesia.	• Fuzzy Logic • Artificial Neural Network • K-Nearest Neighbors	• Mean Absolute Percentage Error (MAPE) **Validation:** 18 training observations	• The fuzzy Logic method produces the best accuracy with a MAPE value of 88%.	• FL method has high accuracy due to no training process involved in the method modeling. • The studies used 18 training samples that are too small and only applied to ANN and KNN algorithms that heavily affected the training process. • A comprehensive comparison of predictive models is from the mathematical framework and machine learning. • The performance evaluation using one measurement metric of error. • The sample size used in this study is too small.
[17]	3,527 transactions of residential apartments in Nicosia, Cyprus from the year 2008 to 2014. Collected from the Cyprus Department of Lands and Surveys, and the Central Bank of Cyprus' Residential Index	• Random Forest • Linear Regression	• Linear coefficient, α • Root Mean Square Error (RMSE) • Mean Absolute Error (MAE) • Mean Absolute Percentage Error (MAPE) • Average Sales Ratio (SR) **Validation:** 80:20 train-test split ratio	• Random Forest has better fitted the model than Linear Regression with the means of the differences of 9.73% for the Linear coefficient. • Random Forests, −1.27% for the RMSE, 1.44% and 2.07% for the MAE and MAPE, and −0.73% for the SR.	• Most of the variables have been excluded in the data pre-processing stage. • The learning algorithm compares information-based learning and error-based learning. • A comprehensive comparison of the predictive model performances using five error metrics measurements. • The sample size used in this study is big.
[18]	2,462 rental listing in Beijing, China from November 2016 to January 2017. Obtained from 58.com website.	• Gradient Boosting Regression Tree • Linear Regression • Regression Tree • Random Forest	• Root Mean Square Error (RMSE) • Correlation coefficient, r **Validation:** 70:30 train-test split ratio	• Negative correlation between Rent and Distance from the City Centre. • The Tree-based models outperformed the Linear Regression Model, with the Regression Tree has a correlation coefficient of 0.57 in the test set and the smallest RMSE value of 1.05.	• The data are fundamentally noisy. • The learning algorithm compares information-based learning and error-based learning. • The performance evaluation using two measurement metrics of error. • The sample size used in this study is big.

Article, Author, and Year	Sample Data	Algorithm	Validation Method and Evaluation Measurement	Result	Strength/ Limitation
[19]	5,000 samples of real estate price quarterly data in Mumbai from the year 2005 to 2016. Collected using web scraping from websites like 99acres.com, Magicbricks.com, and Google.com.	• K-Nearest Neighbors • Random Forest • Linear Regression • Linear Regression (Gradient Descent)	• Root Mean Square Error (RMSE) • Mean Absolute Percentage Error (MAPE) • Mean Absolute Error (MAE) **Validation:** 80:20 train-test split ratio	• Random Forest algorithm was found to have less error with an RSME value of 0.007, MAE value of 0.063, and MAPE value of 6.328%.	• The variables used in this study are based on house location factors. • A comprehensive comparison of predictive models is from information-based learning, similarity-based learning, and error-based learning. • The performance evaluation using three measurement metrics of error. • The sample size used in this study is big.
[20]	21,613 entries of housing sales in King County, USA. Collected from the Kaggle website.	• Support Vector Machine • Linear Regression	• R-squared • Mean Absolute Error (MAE) • Mean Square Error (MSE) • Root Mean Square Error (RMSE) **Validation:** 80:20 train-test split ratio	• There is no difference between the performance of feature selections and feature extraction. • Both achieve 0.86 R-squared scores after log transformation on house prices. • The best combination of parameters that achieves the highest R-squared score is SVR with RBF kernel and C sets to 10.	• Applying several methods for feature extraction, and feature selection to reduce the high dimensional dataset. • No comparison between learning where this study only applying error-based learning. • The performance evaluation using four measurement metrics of error. • The sample size used in this study is big.
[21]	16,472 price records for new housing units in Santiago, Chile.	• Linear Regression • Neural Network • Support Vector Machine • Random Forest	• Root Mean Square Error (RMSE) • Mean Absolute Error (MAE) • R-squared **Validation:** 70:30 train-test split ratio	• The RF algorithm outperformed the LR, NN and SVM for training dataset (R^2 = 0.98, RMSE = 0.019, MAE = 0.00), and validation dataset (R^2 = 0.956, RMSE = 0.041, MAE = 0.003).	• The dataset used in this study is based on the new housing unit which consists of apartments and houses. • A learning comparison of predictive models is from the information-based learning, and error-based learning. • The performance evaluation using three measurement metrics of error. • The sample size used in this study is big.

3. Methodology

This is an applied study using quantitative Penang's terrace houses dataset for the years 2018 and 2019 collected from brickz.my. This study involves the development of three ML techniques namely Multiple Linear Regression (MLR), Random Forest (RF), and K-Nearest Neighbors (KNN). The dataset is trained using these three ML techniques to predict Penang's terrace house prices.

3.1 Operational Framework

Figure 3.1 shows the illustration of the operational framework which is adapted into this study. The framework includes four main phases which are: Phase one; data acquisition and data pre-processing, Phase two; exploratory data analysis, phase three; model development and performance evaluation, and finally, Phase four; model implementation and conclusion.

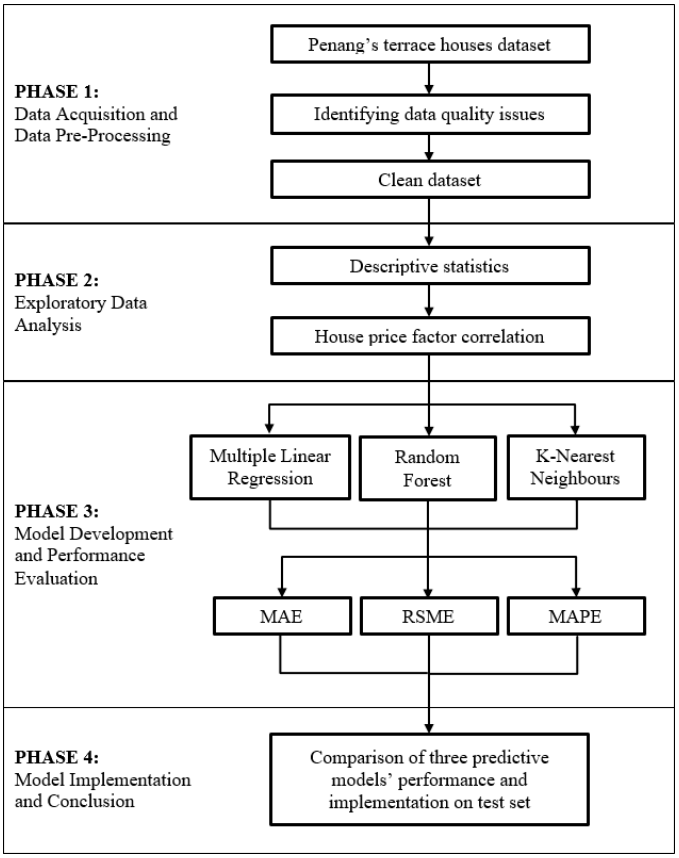

Figure 3.1 Operational Framework

To begin the data analysis, the data in raw form are not ideal. To be beneficial for ML, the data must first be cleaned from missing data, wrong spelling, duplicates, and outliers. The dataset undergoes the cleaning process includes renaming the heading, creating new re-coded data for categorical features via one-hot encoding, regrouping the data, removing uncommon features, and changing data type. This process is used to get the optimal number of features that describe all the important information in the dataset [22]. The pre-processing task is done for Penang's terrace houses dataset in Jupyter Notebook and using matplotlib libraries.

The process involved in phase two is exploratory data analysis. This phase comprises two main activities, namely data exploration and house price factor correlation. In this phase, the dataset undergoes two types of analysis, namely the descriptive analysis (also called the explanatory data analysis) and correlation analysis. The task is done for Penang's terrace houses dataset in Jupyter Notebook and using Seaborn as well as matplotlib libraries.

The explanatory data analysis or graphical exploration goals are to fully understand and to provide an in-depth preliminary investigation on the characteristics of the dataset and to determine any data quality issues that exist in the dataset. In particular, this process is to identify outliers, examine, and descriptive statistics of all the features [23].

The correlation analysis is one way to measure the strength of the relationship between two continuous or numerical features. A descriptive feature that correlates strongly with the house price (target feature) would be a good place to start building a predictive model. By determining which input features are associated with the house price will ensure that only relevant features are included in the model. Consequently, to produce a fitted house price prediction model. This analysis is done on all the features using the Pearson correlation coefficient, represented by the r-value.

The clean dataset will then be used to fit three ML models, and later will be evaluated using several performance measures in phase three. This procedure is achieved by using the Alteryx Designer 2020.4. Figure 3.2 illustrates the general process of model development and performance evaluation.

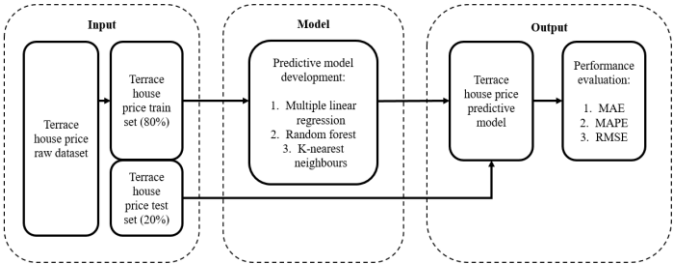

Figure 3.2 Model Development and Evaluation

After the performance evaluation of the three ML models is done, the selected model will be implemented on the test set of Penang's terrace houses dataset using the Alteryx Designer 2020.4. From the implementation, the prediction result is observed. The visualization of the results is performed using Tableau Desktop Professional Edition 2020.3.1.

3.2 Data description

This study collects Penang's terrace houses dataset from brickz.my (also called the Brickz). Brickz is an independent property website. The dataset consists of actual property transactions. The source of data in Brickz is from the VPSD which officially records a property transaction once the stamp duty for the Sales and Purchase is paid. Brickz has been compiling these officially recorded transactions since January 2014 and will be updating the transacted data monthly [24].

The collected dataset consists of 2,699 sold terrace houses in Penang, Malaysia with nine features or variables. The dataset is the transaction of the sold terrace houses in Penang from January 2018 until December 2019. The dataset contains nine features with three categorical data types and six are numerical data types. Table 3.1 summarizes the dataset features that are provided by Brickz.

Table 3.1 Features described in the dataset

Feature Name	**Description**	**Data Type**
Location	The property area or district *(consists of 33 areas)*	Categorical
Building_Type	The property building type *(intermediate, corner lot, or end lot)*	Categorical
Tenure	The property tenure type *(freehold – 1, least hold – 0)*	Categorical
Floors	The number of floors of the property *(1, 1.5, 2, 2.5, 3, 3.5)*	Numerical
Rooms	The number of rooms of the property *(0, 1, 2, 3, 4, 5, 6, 7)*	Numerical
Land_Area	The size of the property land area	Numerical
Build_Up	The size of the property	Numerical
Price_Psf	The property price per square feet	Numerical
Price	The property sold price	Numerical

4. Result and Discussion

4.1 Data pre-processing and exploratory data analysis

In data pre-processing, the dataset undergoes the cleaning process that includes removing missing or duplicate information, filtering meaningless data, and consolidating distinct data representations to have consistent and accurate datasets. Followed by renaming the heading, encoded the categorical features as a number, regrouping the data, removing uncommon features, and changing the data type to ensure the dataset is complete and accurate for model development.

Generally, ML algorithms require data in numerical form, especially for Single and MLR, although some of them natively use categorical variables or features [25]. Thus, the categorical variable must be encoded as numbers (one number per category) for the MLR model development [26]. One of the approaches is by one-hot encoding method where each of the categories for the categorical feature is encoded into a separate binary variable that has a value of ‘1’ and ‘0’. Sometimes the variable created using this method are called ‘dummy variables’ [25].

As described in Table 3.1, the dataset contains three features with the categorical data type, namely ‘Location’, ‘Building_Type’, and ‘Tenure’. For ‘Building_Type’, the category for the feature is replaced with three separate binary features, namely ‘Building_Type_INTERMEDIATE’, ‘Building_Type_CORNER LOT’, and ‘Building_Type_END LOT’ as new features in the dataset.

The ‘Location’ feature, contains 33 location categories, which leads to the creation of 33 new variables in the dataset. Due to that, the current ‘Location’ feature is divided into five separate groups based on the districts in Penang, namely South Seberang Perai (SP), Central Seberang Perai (CP), North Seberang Perai (NP), Northeast Penang Island (NE), and Southwest Penang Island (SW) [27]. Thus, the

new five features are introduced as 'Location_SP', 'Location_CP', 'Location_NP', 'Location_NE', and 'Location_SW' as the representation of five districts of Penang.

As for the 'Tenure' feature, no encoded required because the categories already have a value of '1', and '0', which is for freehold and least hold. With the creation of an additional eight new features via one-hot encoding, the total features that will be used for the model development are 15 features. Table 4.1 describes all the features after the data pre-processing procedure. Provided in Figure 4.1, the summarization of the outcome for the data pre-processing task.

Table 4.1 Features description after data pre-processing task

Feature Name	**Description**
Price	The property sold price
Tenure	The property tenure type *(freehold – 1, least hold – 0)*
Floors	The number of floors of the property *(1, 1.5, 2, 2.5, 3, 3.5)*
Rooms	The number of rooms of the property *(0, 1, 2, 3, 4, 5, 6, 7)*
Land_Area	The size of the property land area
Build_Up	The size of the property
Price_Psf	The property price per square feet
Building_Type _CORNER LOT	The property with corner lot building type *(corner lot = 1, not corner lot = 0)*
Building_Type _END LOT	The property with end lot building type *(end lot = 1, not end lot = 0)*
Building_Type _INTERMEDIATE	The property with intermediate building type *(intermediate = 1, not intermediate = 0)*
Location_CP	The property located at Central Seberang Perai *(if yes = 1, if no = 0)*
Location_NE	The property located at Northeast Island *(if yes = 1, if no = 0)*
Location_NP	The property located at North Seberang Perai *(if yes = 1, if no = 0)*
Location_SP	The property located at South Seberang Perai *(if yes = 1, if no = 0)*
Location_SW	The property located at Southwest Island *(if yes = 1, if no = 0)*

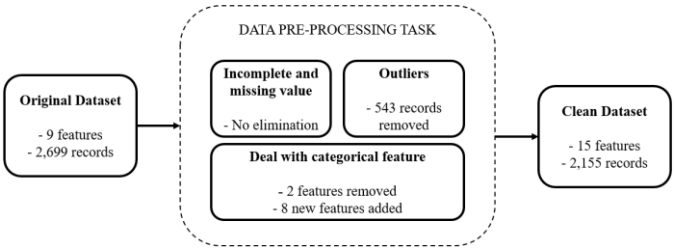

Figure 4.1 The outcome of the Data Pre-processing Task

After the data pre-processing task conducted earlier has resulted in a clean dataset consists of 2,155 records. To gain a preliminary understanding of the dataset, the data analysis is conducted to see the data pattern and distribution. All 2,155 records have undergone two types of data analysis, namely the descriptive statistic (also called the explanatory data analysis) and correlation analysis as well as a feature selection procedure.

4.1.1 Descriptive statistics

Table 4.2 summarizes the descriptive statistic for all the dataset features for the numerical variables which are 'Price', 'Floors', 'Rooms', 'Land_Area', 'Built_Up', and 'Price_Psf' which is comprising of Mean, Standard Deviation, Minimum, Quartile 1 (25%), Quartile 2 (50%), Quartile 3 (75%), and Maximum.

Table 4.2 Descriptive statistic for the dataset

Feature	Mean	SD	Min	25%	50%	75%	Max
Price	420,270.50	260,392.80	20,000.00	232,000.00	363,000.00	520,000.00	1,700,000.00
Floors	1.72	0.63	1.00	1.00	2.00	2.00	3.50
Rooms	3.45	0.68	2.00	3.00	3.00	4.00	5.00
Land_ Area	1,336.35	238.20	700.00	1,195.00	1,302.00	1,442.00	2,055.00
Built_Up	1,353.66	561.89	418.00	800.50	1,433.00	1,711.00	2,964.00
Price_Psf	303.34	106.12	27.00	235.00	286.00	356.00	632.00

In summary, the lowest terrace house price is RM 20,000.00 and the most expensive terrace house is sold at RM 1,700,000.00. The average price of the house is RM 420,270.50 which is quite high. As for the land area size, the average size of the land area that has been sold is 1,336.35 sqft and the maximum land size that has been sold is 2,055.00 sqft. Meanwhile, the average built-up area is 1,353.66 sqft and the maximum built-up size is 2,964.00 sqft which indicates the terrace houses in Penang are huge.

4.1.2 Correlation analysis

This analysis is done on all the features using the Pearson correlation coefficient represented by the r-value. Table 4.3 shows the Pearson correlation coefficient, r value between the numerical features, and the house price. Figure 4.2 visualizes the correlation between the numerical features and the house price in an image of a heatmap.

Table 4.3 Correlation between variables and 'Price' using the Pearson correlation coefficient

Feature	Pearson correlation coefficient, r value
Built_Up	0.775
Price_Psf	0.705
Floors	0.703
Rooms	0.560
Land_Area	0.385

The correlation analysis using the Pearson correlation coefficient represented by r value has identified four variables or house features that highly correlated with the house price; namely the size of the property ('Built_Up'), the property price per square feet ('Price_Psf'), the number of floors of the property ('Floors'), the number of rooms of the property ('Rooms').

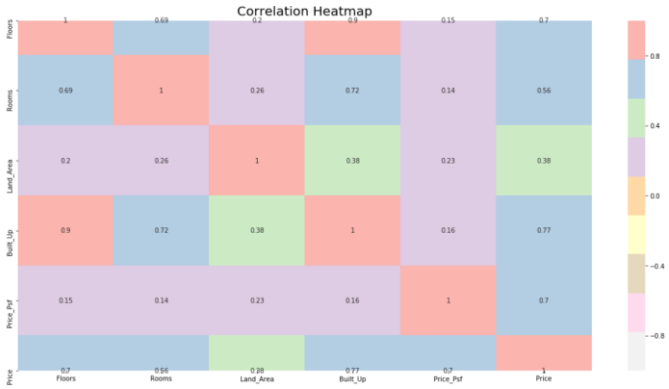

Figure 4.2 Correlation Heatmap between House Features

The heatmap also reveals the multicollinearity that exists between 'Built_Up', and 'Floors' with an r-value of 0.9. Multicollinearity is a case where multiple independent variables are highly correlated with each other [28] and are highly recommended to be resolved before starting the process of model development [29].

4.1.3 Feature selection

In the data-preprocessing task, the categorical variables undergo an encoding process to become new numerical variables using the one-hot encoding method. However, the one-hot encoding variables are highly correlated or multicollinearity [30]. As stated by [28] and [29], multicollinearity is a case where multiple predictor variables are highly correlated with each other. These studies also provide several remedies to fight multicollinearity such as variable elimination, increase the sample size, applying ridge regression, and principal component regression. Both authors conclude that the widely used approach as a multicollinearity remedy is variable elimination. Hence, for this house price prediction model, the independent variables that are highly correlated with each other are eliminated from the dataset.

Referring to Table 4.1, the one-hot encoding variables are 'Building_Type_CORNER LOT', 'Building_Type_END LOT', and 'Building_Type_INTERMEDIATE' created from the 'Building_Type' variable. The 'Location_SP', 'Location_CP', 'Location_NP', 'Location_NE', and 'Location_SW' are created from the 'Location' variable. Consequently, introduces multicollinearity. To reduce the correlation among variables, one variable from the one-hot encoded variable cannot be used or remove [30] [31]. The eliminated one-hot encoding variables are 'Building_Type_CORNER LOT' and 'Location_CP'. Similarly, from Figure 4.2, the correlation heatmap reveals there are multicollinearity issues found between 'Built_Up', and 'Floors' with an r-value of 0.9. The eliminated features are the 'Floors' features.

The different set of feature groups is summarized in Table 4.4. In this study, two groups are used for ML predictive model development. The first group contains all features dataset and the second group contains selected features dataset. The selected features dataset contains 11 features where the eliminated variables where the independent variables are highly correlated with each other from the original dataset, to avoid data redundancy. Hence, to capture the different performances from both groups.

Table 4.4 Different feature selection of the dataset

Feature	All Features	Selected features
Built_Up	√	√
Price_Psf	√	√
Floors	√	Not included
Rooms	√	√
Land_Area	√	√
Tenure	√	√
Location_NP	√	√
Location_NE	√	√
Location_SW	√	√
Location_SP	√	√
Location_CP	√	Not included
Building_Type_INTERMEDIATE	√	√
Building_Type_END LOT	√	√
Building_Type_CORNER LOT	√	Not included

4.2 Model Development and Performance Evaluation

4.2.1 Multiple Linear Regression

In Alteryx Designer, developing the Multiple Linear Regression (MLR) predictive model requires several parameter settings such as target variable selection and predictor variables selection [32]. For MLR predictive model development, the target variable is the house price ('Price'). As for the predictor variables, two different feature selection group are used; all feature group which consists of 14 variables and selected features group contains 11 variables.

For MLR model developments, Alteryx Designer provides the customize parameter for validation purposes such as omit model constant, use weight variable for weighted least square, and use regularized regression. Also, cross-validation to determine estimates of model quality is a customize option to modify the model settings [32]. However, for this study, the customization parameters are set as default.

The Alteryx workflow for both feature selection datasets is illustrated in Figure 4.3 and Figure 4.4, respectively. Referring to the workflow in Figure 4.3, the model development started by uploading Penang's house prices dataset with all the house features dataset into the Alteryx Designer canvas. The dataset is normalized between 0 and 1 using the Normalize Columns tool. Then, in the hold-out step, the dataset is split into train-test (estimation-validation) set with 80% train set and 20% test set (80:20) proportion using Create Samples tool. The total number of records for the train set is 1,724 and 431 records for the test set. After that, in model development, all the specifications described are set and feed with the train set. Finally, in data evaluation, the predictive model is validated with a test set using the Model Comparison tool. The binocular at the output shows the report result for both train-test sets that include the MAE, RMSE, MAPE value, and several graphs.

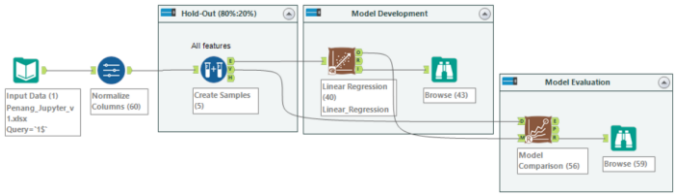

Figure 4.3 Alteryx workflow for MLR predictive model using all features dataset

As for the selected features dataset, the model development workflow is illustrated in Figure 4.4. It is a similar workflow of all features dataset with an additional step after uploading the Penang's terrace house into the Alteryx canvas. It involves selecting the features process in the hold-out step.

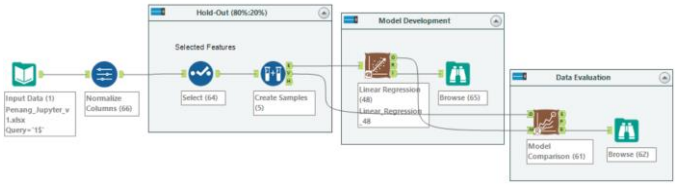

Figure 4.4 Alteryx workflow for MLR predictive model using selected features dataset

4.2.2 Random Forest

Developing a Random Forest (RF) predictive model in Alteryx Designer requires several parameters such as target variable selection, predictor variables selection, number of trees to use for the model, and number of variables to select between at each split [33]. For RF predictive model development, the target variable is the house price ('Price'). As for the predictor variables, two different feature selection group are used; all feature group which consists of 14 variables and selected features group contains 11 variables.

Highlighted by [18], and [34], the output of RF primarily depends on the maximum number of features and the number of the estimator. The maximum number of features parameter is the number of predictors chosen randomly at each tree node. When the parameter is low, the trees become more complex and diverse. However, if the parameter is high (e.g., close to the total number of features), the trees in the forest will tend to be very similar. However, in this study, the number of the variable is set as default.

The number of estimators is one of the most important parameters to control overfitting in RF. It represents the maximum depth or number of trees that can grow in the model. Where increasing the number of trees tends out to be a better solution [18]. However, in this study, the number of trees used, and compared are 100, 250, and 500, representing small, medium, and large trees. Another optional parameter available in Alteryx Designer is set to default value where the minimum five records are allowed in a tree node, as well as 100% records used to create each tree.

The Alteryx workflow for both feature selection datasets is illustrated in Figure 4.5 and Figure 4.6, respectively. Referring to the workflow in Figure 4.5, the model development started by uploading Penang's house prices dataset with all the house features dataset into the Alteryx Designer canvas. Then, in the hold-out step,

the dataset is split into train-test (estimation-validation) set with 80% train set and 20% test set (80:20) proportion using Create Samples tool. The total number of records for the train set is 1,724 and 431 records for the test set. After that, in model development, all the specifications are set and feed with the train set. The three icons of Random Forest (RF) predictive tools represent three models of a different number of trees which are 100, 250, and 500. Finally, in data evaluation, the three predictive models are validated with a test set using the Model Comparison tool. The binocular at the output shows the report result for both the train and test model that includes the MAE, RMSE, MAPE value, and several graphs.

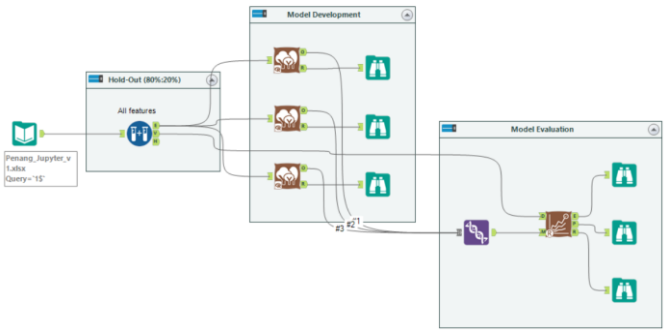

Figure 4.5 Alteryx workflow for RF predictive model using all features dataset

As for the selected features dataset, the model development workflow is illustrated in Figure 4.6. It is a similar workflow of all features dataset with an additional step after uploading the Penang's terrace house into the Alteryx canvas. It involves selecting the features process in the hold-out step.

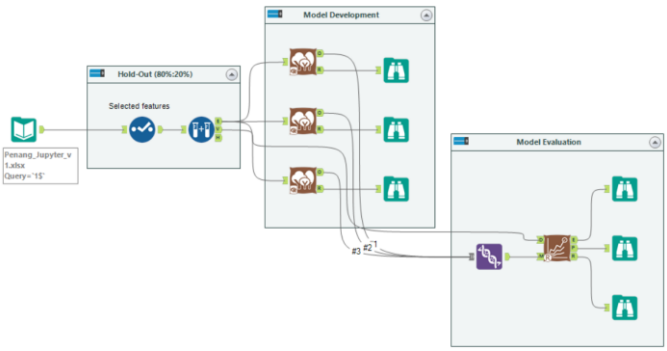

Figure 4.6 Alteryx workflow for RF predictive model using selected feature dataset

4.2.3 K-Nearest Neighbors

In Alteryx Designer, the Find Nearest Neighbors tool is used to develop the K-Nearest Neighbors (KNN) predictive model. It requires several parameters such as a unique key field, and fields [35]. For KNN predictive model development, the unique key field is the house price ('Price'). As for the fields, two different feature selection group are used; all feature group which consists of 14 variables and selected features group contains 11 variables. Other than that, fields standardize option; z-score standardization, or unit-interval standardization, number of neighbors, and the algorithm option to find the nearest neighbors parameter are available in Alteryx Designer [35].

According to [10], [16], [19], [30], and [36], the implementation of KNN is to calculate the average distance of the numerical target of its closest *k* matches in the dataset. However, when calculating the distance, it is necessary to normalize dataset variables or features to avoid the influence of large value features on smaller value features. Thus, for continuous variables, the min-max normalization, or z-score standardization can be used to normalize the values of the features [36]. Therefore, in this project, the z-score standardization is used for feature normalization.

Another parameter that needs to be determined in the Alteryx Designer is the number of neighbors. According to [31], the advantage of choosing *k* bigger than 1 is that higher values of *k* can reduce the risk of overfitting due to noise in the training data. If the *k* is too low, the model may be fitted with the noise in the data. However, if the *k* is too high, the model might not capture the local structure in the data. Then, in this study, the *k* range that has been set, and compared is 1, 3, and 6 representing small, medium, and a large number of neighbors.

The last parameter that needs to be tuned in Alteryx Designer is the algorithm designated to find the nearest neighbors. There are several options of the algorithm provided for finding the nearest neighbors provided in Alteryx Designer such as Cover Tree, KD-Tree, VR, CR, and Linear search [35]. The Cover Tree is designed to facilitate the speed-up of the nearest neighbor's search and developed for indexing low-dimensional data [37]. The KD-Tree is useful for searches involving multidimensional search keys and is a special case of binary space portioning trees [38]. Then, VR is the method used by Venables and Ripley (2002). CR is a version of the VR algorithm based on a modified distance measure, and Linear Search is an algorithm that calculates the distance between each point in the query stream to all the points in the data stream [35]. Therefore, in this study, the algorithm designated to find the nearest neighbors is Linear Search.

The Alteryx workflow for both feature selection datasets is illustrated in Figure 4.7 and Figure 4.8, respectively. Referring to the workflow in Figure 4.7, the model development started by uploading Penang's house prices dataset with all the house features dataset into the Alteryx Designer canvas. Then, in the hold-out step, the dataset is split into train-test (estimation-validation) set with 80% train set and 20% test set (80:20) proportion using Create Samples tool. The total number of records for the train set is 1,724 and 431 records for the test set. After that, in model development, all the specifications are set and feed with the train set. Finally, in data evaluation, the predictive models are validated with a test set using the Find Nearest Neighbors tool. The results for KNN are produced using the Formula tool, and the binocular at the output shows the performance result for the predictive model based on the RMSE.

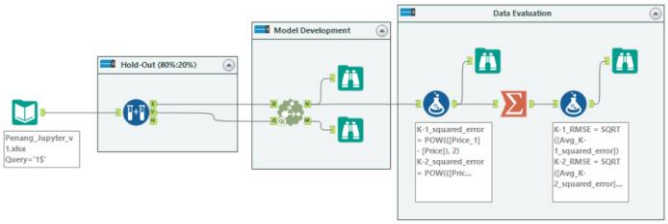

Figure 4.7 Alteryx workflow for KNN predictive model using all feature dataset

As for the selected features dataset, the model development workflow is illustrated in Figure 4.8. It is a similar workflow of all features dataset with an additional step after uploading the Penang's terrace house into the Alteryx canvas. It involves selecting the features process in the hold-out step.

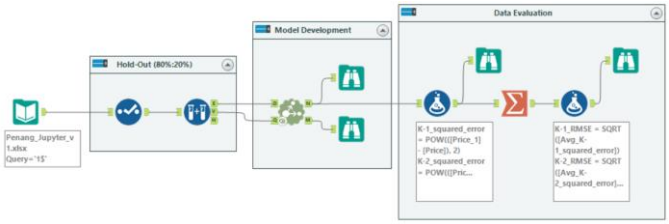

Figure 4.8 Alteryx workflow for KNN predictive model using selected feature dataset

4.2.4 Model Evaluation, Comparison, and Selection

Table 4.5 shows the comparison of MLR predictive models between two different feature selection groups; all features and selected features. In general, for the train set, the predictive model using all features dataset has fitted the model better than the predictive model using the selected features dataset, although the differences are small. Further observation, in the model validation using the test set, the predictive model with selected features dataset has produced lower prediction error compared to the predictive model with all features dataset. Although, the differences are small. It reveals that, even though the model using all features dataset has better fit the model using train set, it produced more prediction error when applying the test set. On the other hand, the model utilizing the selected features dataset has fitted the model using the train set and produced a slightly lower prediction error.

However, the benchmark used to evaluate the prediction model performance is provided by the lower prediction error produced by the model in the test set [40]. Due to that, the predictive model performance remained unbiased and efficient using the selected features dataset with the performance measurement result of 53,965.094 for RMSE, 36,095.976 for MAE, and 14.4% MAPE for the train set. As for the test set, the RMSE value is 58,679.909, for MAE is 38,712.734, and 12.9% for MAPE for the best performance accuracy.

Table 4.5 Performance measurement comparison for MLR models

Train Set (80%)				**Test Set (20%)**		
Feature Selection	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
All Features	53,910.257	35,995.753	14.4%	58,719.533	38,704.377	12.9%
Selected features	53,965.094	36,095.976	14.4%	58,679.909	38,712.734	12.9%

Next, the comparison among RF models using different parameters and feature selection datasets are shown in Table 4.6. In general, the predictive model using all features dataset has fitted the model better compared to the predictive model using selected features dataset in the train set. Similarly, the predictive model using all features dataset has produced lower prediction error in the model validation with the test set compared to the predictive model using the selected features dataset. The MAPE differences are around 1% - 2% for both groups. Further observation, in terms of the number of trees comparison in each predictive model, the performance result

differences are small in the train set. Similarly, there are small differences in the prediction error produced in model validation using the test for the individual model. Also, looking into the number of trees comparison between two groups, the predictive model utilizing 250 number of trees fitted the model better and produced lower prediction error compared to the model utilizing 100 and 500 number of trees. It reveals that 250 trees are enough samples to reduce the bias-ness of the data. Beyond that, it will increase the bias of the result.

In summary, the predictive model utilizing all features dataset with 250 trees achieve both highest performance for train and test set, with 23,786.856 for RMSE, 13,769.965 for MAE, and 4.674% MAPE for the train set. As for the test set, the RMSE value is 35,612.956, for MAE is 20,816.257, and 6.096% for MAPE for the best performance accuracy.

Table 4.6 Performance measurement comparison for RF models

Train Set (80%)				**Test Set (20%)**		
All Features Dataset						
Number of Tree	RMSE	MAE	MAPE	RMSE	MAE	MAPE
100	24,495.353	14,271.092	4.710%	36,869.849	21,687.347	6.314%
250	23,786.856	13,769.965	4.674%	35,612.956	20,816.257	6.096%
500	24,020.341	13,696.290	4.703%	36,840.454	21,276.859	6.179%
Selected Features Dataset						
Number of Tree	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
100	35,310.029	21,173.374	6.934%	49,775.273	29,005.077	8.173%
250	32,220.969	19,912.295	6.602%	45,832.561	27,135.134	7.960%
500	32,786.556	20,122.848	6.620%	46,543.119	27,458.216	7.952%

Finally, the comparison among KNN algorithms using different parameters and feature selection datasets are shown in Table 4.7. In general, the predictive models that utilized all features dataset perform better than the models utilizing the selected features dataset. Also, from the table, the result shows that as the number of neighbors increases, the value of RMSE is increases indicates the model performance is reduced for the predictive models utilizing both feature selection groups. Further observation, in terms of the number of neighbors, *k* comparison in each predictive model, predictive model using selected features dataset with $k = 1$ has produced better performance compared to other predictive models. Probably because the data is scattered, and nearest neighbors are often fairly distant [25]. In summary, the selected features dataset with a number of the nearest neighbor of 1 ($k = 1$) is the best performance accuracy model for KNN with the lowest RMSE value of 59,737.252.

Table 4.7 Performance measurement comparison for KNN models

Number of Nearest Neighbors	**RMSE**
All Features Dataset	
1	63,073.701
3	97,473.000
6	113,884.047
Selected Features Dataset	
1	59,737.252

Number of Nearest Neighbors	RMSE
3	98,056.683
6	124,120.626

Overall, for all models, the selected features dataset produces better performance results compared to the all features dataset. However, in terms of model parameters, tuning the value for RF and KNN has produced minor differences in each set of the dataset. Comparing all three algorithms, MLR, RF, and KNN, it shows that the RF model trained using all features dataset with a parameter of 250 number of trees is the most suitable model to be used in predicting the house prices in Penang. It shows that the RF model trained using all features dataset with a parameter of 250 number of trees has better fitted the model compared to other models. Moreover, when the model is validated with the test set, RF using all features dataset with a parameter of 250 number of trees has produced the lowest error.

4.3 Model Implementation

The best ML predictive model has been selected and used to predict the test set (20% from total records) of Penang's terrace house price dataset. The test set consists of 431 Penang's terrace house prices. Figure 4.9 shows the Alteryx Designer workflow to execute the process which involves several steps. First, the RF model with 250 trees using all features dataset is trained. Next, the model is connected to the Score tool to predict the test set. Finally, the score is extracted into Tableau Hyper Data Extract (.hyper) for visualization.

The house price prediction result that has been produced by the Alteryx workflow in Figure 4.9 is fed into Tableau Desktop Professional Edition 2020.3.1, a visualization software to illustrate the descriptive analysis of the prediction and to compare with the test set of Penang's terrace house price dataset.

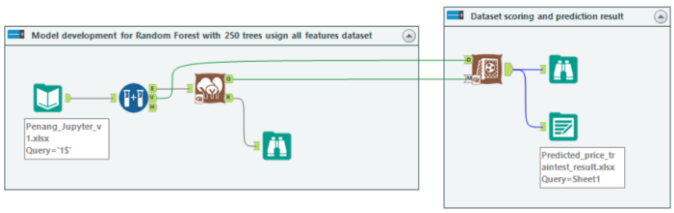

Figure 4.9 Alteryx workflow for RF with 250 trees using all feature dataset to predict the house prices

The actual price and predicted price are plotted as in Figure 4.10. The scatterplot shows that there is a linear trend between the actual price and predicted price for Penang's terrace house dataset. Thus, indicates the RF with 250 trees using all feature dataset performs well compared with the other models.

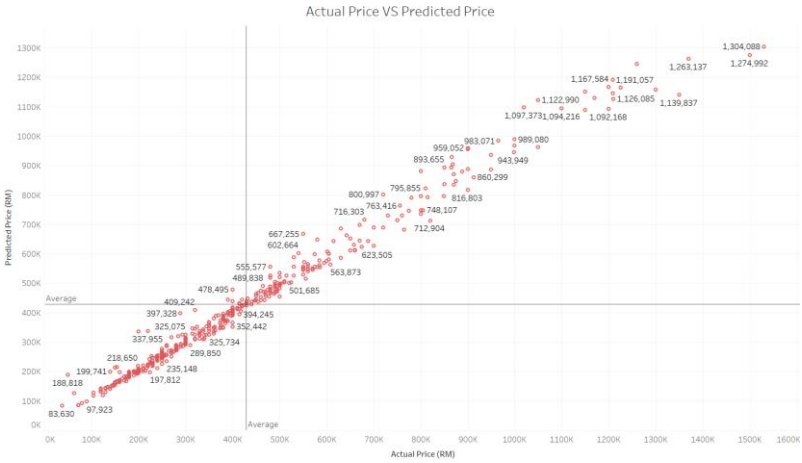

Figure 4.10 Actual price versus predicted price

The column chart in Figure 4.11 shows the comparison of average house value between the actual price and predicted price in Penang's terrace house price dataset. The average house value for the actual price is RM 428,800.00, and the average house value for the predicted price is RM 427,415.00. Comparing to the average actual price, the average house value for the predicted price is slightly lower. Thus, this indicates the model tends to predict a lower price than the actual price.

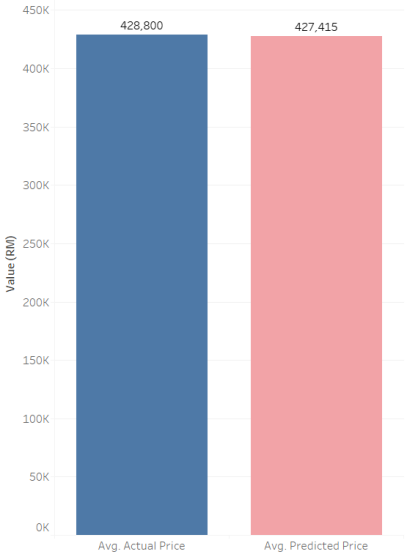

Figure 4.11 Comparison of the average actual price and average predicted price

The next following charts illustrate the comparison of the average actual price and average predicted price for Penang's terrace houses based on the house features which consist of building type, house location, tenure type, number of floors, and number of rooms.

For the average price comparison between building types, Figure 4.12 shows that End Lot terrace houses have the highest average price for both actual and predicted prices with RM 485,160.00 and RM 475,916.00, respectively. Furthermore, the figure reveals that the Intermediate type has the lowest average price for both actual and predicted prices with RM 424,631.00 and RM 423,654.00, individually. On top of that, for the Corner Lot type, the average actual price is a

little lower than the predicted price, with RM 453,000.00 and RM 455,111.00, respectively. In general, the average house value for the predicted price is slightly lower for the End Lot and Intermediate type, and higher for the Corner Lot type.

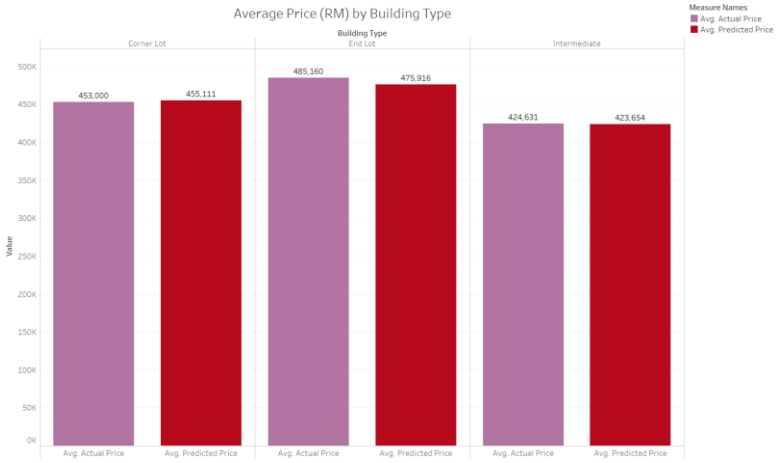

Figure 4.12 Comparison of the average actual price and average predicted price by building types

From Figure 4.13, based on the house location, Northeast Penang Island (NE) have the highest average price for both actual and predicted price with a house value of RM 908,677.00 and RM 917,725.00, respectively. It is not surprising, as NE is situated within the heart of George Town, which is also Penang's capital city [41]. This leads to expensive house prices. The second highest average price for Penang's terrace houses is at Southwest Penang Island (SW) with RM 774,405.00 for actual price and RM 761,888.00 for predicted price. Next, North Seberang Perai (NP), with an average value of RM 409,430.00 for the actual price and RM 406,784.00 for the predicted price. Followed by Central Seberang Perai (CP) with an average actual price of RM 373,595.00, and an average predicted price of RM 372,243.00. The lowest average house value is at South Seberang Perai (SP) for both actual and predicted prices with the house value of RM 284,940.00, and RM 285,403.00, individually. In general, the average house value for the predicted price is slightly lower for houses located at SW, CP, and NP. However, the average value for NE and SP is predicted higher than the actual price.

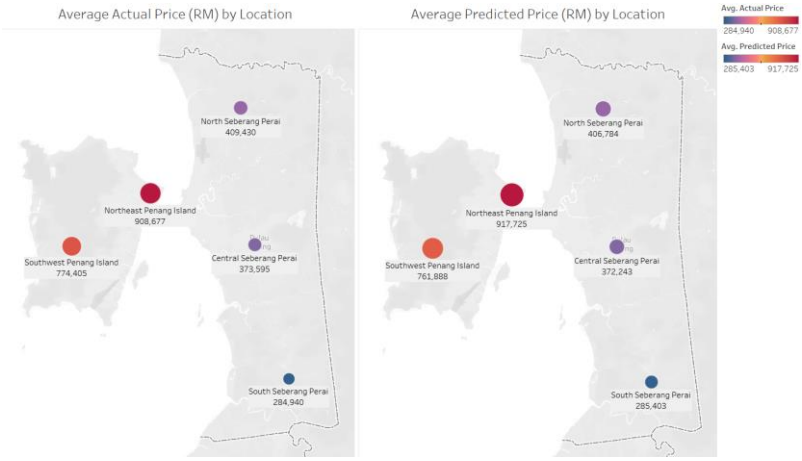

Figure 4.13 Comparison of the average actual price and average predicted price by terrace house location

Figure 4.14 illustrates the comparison of the average house price by tenure type in Penang. The result shows that the average actual price for freehold houses in Penang is RM 435,405.00, and the predictive model gives the average predicted price of RM 433,570.00 for the same tenure type. On the other hand, the leasehold average actual price is RM 311,636.00, and the average predicted price is RM 318.225.00. In general, terrace houses with freehold tenure types are much more expensive than the leasehold tenure type in Penang.

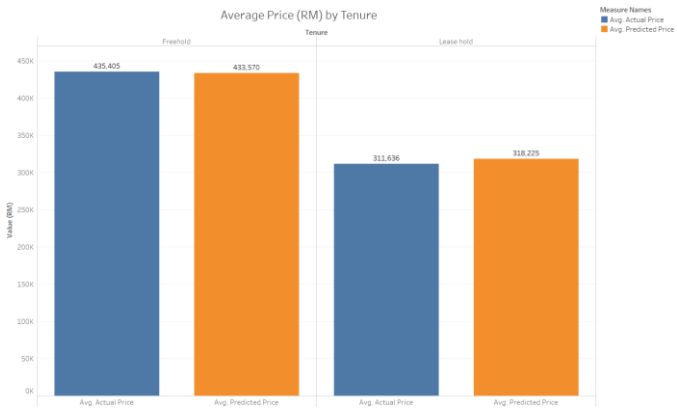

Figure 4.14 Comparison of the average actual price and average predicted price by tenure type

The result in Figure 4.15 shows the comparison of the average house price by the number of floors in Penang. The number of floors for the terrace houses includes houses with only one floor, two floors, two and a half floors, three floors, and three and a half floors. In general, the graph shows the house price increase with an increase in the number of floors. The most expensive houses in Penang with an average actual price of RM 1,179,000.00 and an average predicted value of RM 1,171,175.00 are houses with 3.5 floors. As the houses with 1 floor are the cheapest with an average actual price of RM 242,004.00 and RM 242,905.00 for an average predicted price. Other than that, the average actual price is higher than the average predicted price for houses with two and a half floors and higher, and the average actual price is lower than the average predicted price for houses build with one and two floors. Overall, there are small differences between the average actual price value and average predicted price value for all number floors.

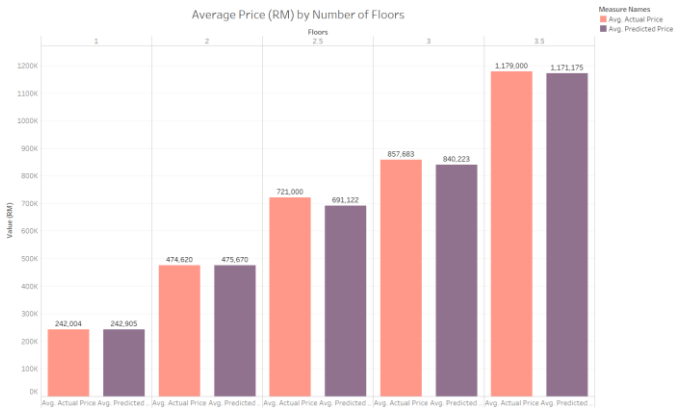

Figure 4.15 Comparison of the average actual price and average predicted price by the number of floors

Figure 4.16 visualizes the comparison of the average actual price and average predicted price by the number of rooms for the terrace houses in Penang. The number of rooms for Penang's terrace houses starts with the minimum number of two, three, four, and five as the maximum number of rooms. Similar to the result by the number of floors, the graph shows the house price increase with an increase in the number of rooms. The most costly houses in Penang with an average actual price of RM 873,986.00 and an average predicted value of RM 850,586.00 are houses with five rooms. As the houses with two rooms are the lowest with an average actual price of RM 189,889.00 and RM 197,872.00 for an average predicted price. In general, the average actual price is higher than the average predicted price for houses with rooms of four and five, and the average actual price is lower than the average predicted price for houses build with two and three rooms. Overall, there are small differences between the average actual price value and average predicted price value for all numbers of rooms.

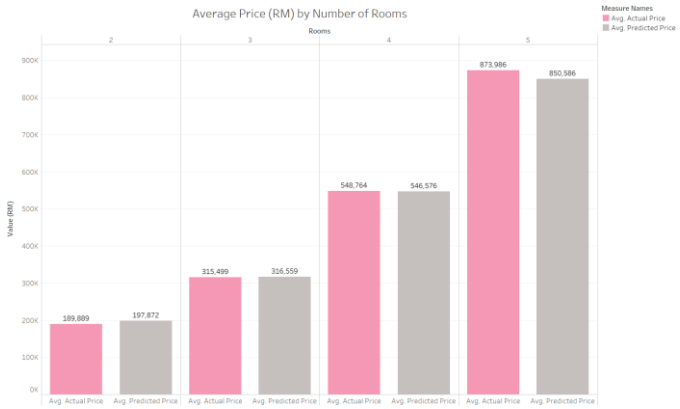

Figure 4.16 Comparison of the average actual price and average predicted price by the number of rooms

The next few maps visualize the comparison of an average actual price and average predicted price based on the house built-up size, and the land area size for the actual and predicted price as per terrace house location.

Figure 4.17 illustrates the average actual and average predicted prices for Penang's terrace houses are based on the built-up size. The biggest average Penang's terrace house built-up size is 1,945.8 square feet at NE. Thus, this location also is the most expensive area with an average actual and predicted price of RM 908,677.00 and RM 917,725.00, respectively. The smallest built-up size and cheapest terrace house in Penang are at SP with an average built-up size of 1,134.9 square feet and average actual and predicted price of RM 284.940.00 and RM 285,403.00, respectively. Furthermore, the average actual price is higher than the average predicted price for houses located at SW, NP, as well as CP, and the average actual price is lower than the average predicted price for houses located at NE and SP. Overall, there are small differences between the average actual price value and average predicted price value for average built-up size at all districts in Penang. In general, the terrace houses located in Penang Island are bigger on the built-up and more costly than the terrace houses located on the mainland.

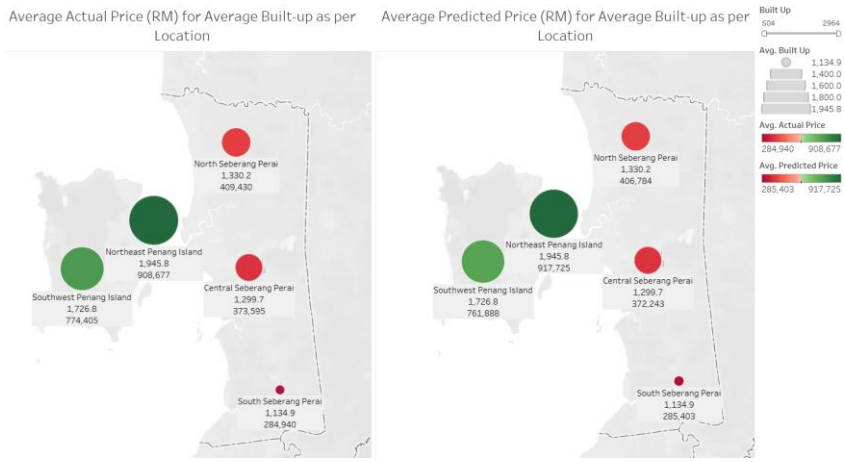

Figure 4.17 The average actual and average predicted price based on the built-up size as per location

In Figure 4.18, the average actual and average predicted prices for Penang's terrace houses are based on the land area size are revealed. The biggest average Penang's terrace house land area size is 1,464.4 square feet at SW. However, this location's average actual and predicted price is RM 774,405.00 and RM 761,888.00, respectively which is not the highest. The smallest average land area size is located at SP. Furthermore, the average land area size between NE, NP, and CP are almost the same, but based on the house value, NE has the highest and most expensive value. Overall, the terrace houses located in Penang Island are bigger in land size area and more costly than the terrace houses located on the mainland.

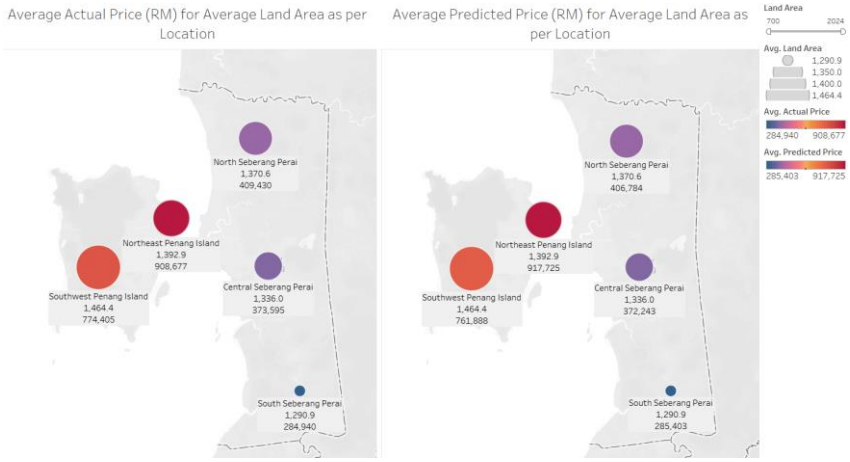

Figure 4.18 The average actual and average predicted price based on the land area size as per location

The visualization in Figure 4.19 is the dashboard produced in Tableau which collects all the house feature discussed in this section. The purpose of the dashboard is to show the descriptive analysis of Penang's terrace house prediction comprehensively includes the different house features as Filters, as well as the comparison between the predictive model outcome and Penang's terrace house price test set.

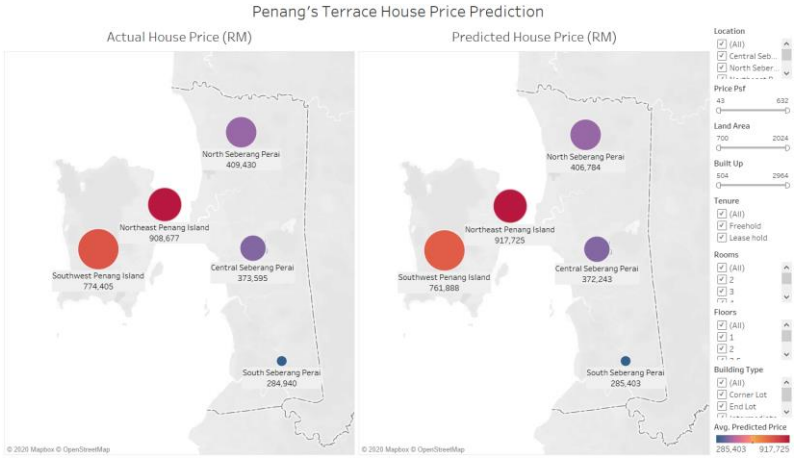

Figure 4.19 Dashboard for Penang's terrace house price prediction using the test set

In summary, the RF with 250 trees using all feature dataset models tends to predict a lower price than the actual price for Penang's terrace house dataset. Although the prediction using RF cannot give as accurate results as the actual house price, the results can provide general insight for the different house features as explained earlier. Overall the difference between the actual house price and predicted house price is small, thus indicate that the small value of prediction error produced by the predictive model. It concludes that the prediction model is significant to use as a house price prediction on Penang's terrace houses.

5. Conclusion

This study explained the findings of data pre-processing and exploratory data analysis conducted on Penang's terrace houses. The data pre-processing procedure has resulted in a clean dataset by eliminating the outliers. The additional features are created via one-hot encoding where the categorical features are encoded into a numerical feature for statistical purposes, thus, to ensure the dataset is complete and accurate for data analysis. The clean dataset consists of 2,155 records out of 2,699 sold terrace houses in Penang, Malaysia.

After that, exploratory data analysis is where the data exploratory has been done using descriptive statistics, histograms, and scatter plots. The correlation analysis using the Pearson correlation coefficient represented by r value has identified four variables or house features that highly correlated with the house price; namely the size of the property ('Built_Up'), the property price per square feet ('Price_Psf'), the number of floors of the property ('Floors'), the number of rooms of the property ('Rooms'). In feature selection, two group features selected are created for the predictive model development stage. The first group contains all variables or features and the second group contains selected variables or features. The selected features group consists of 11 variables out of 14 variables which eliminated the highly correlated variables; 'Building_Type_CORNER LOT', 'Location_CP', and 'Floors'.

Also, the procedures and processes involved in the predictive model development, evaluation, and comparison for three ML algorithms which are MLR,

RF, and KNN are explained and described. In the model development phase, several steps involve such as feature selection, parameter specification, and tuning, as well as data normalization. All three algorithms are developed using two sets of the dataset; all features and selected features and were tuned using several parameters. For MLR, no parameter is tuned and the model is compared between the two sets of the dataset. As for RF, the algorithm is developed with 100, 250, and 500 trees. Similarly, the KNN model is developed with *k* nearest neighbors of 1, 3, and 6. In general, the results show that the selected features dataset have a better performance compared to the all features dataset. However, for the model evaluation, the performance measurement used is MAE, RMSE, and MAPE, and reveals that the RF model with 250 trees using all features dataset outperforms other models.

Also, the selected machine learning predictive model implementation which is Random Forest with 250 trees using all features dataset to Penang's terrace house test set is explained. The results illustrate the general insight that shows the overall Penang's terrace house price prediction.

It is recommended that future work include more house attributes such as location (near to major highways, accessible by public transport), and neighborhood characteristics (population density, nearest school, clinics, and shopping location) This would definitely add more insights to factors affecting prices of terrace houses and would eventually provide a better understanding on Malaysia's real estate market. Also, to explore different pre-processing techniques in achieving better model performance, and to apply different feature selection techniques for feature selection, as well as to include the latest or current dataset in developing the house price prediction model.

References

[1] T. Tech Hong, "Home owning motivation in Malaysia," *J. Accounting, Bus. Manag.*, vol. 1, no. 1, pp. 93–112, 2009.

[2] S. Lip Sean and T. Tech Hong, "Factors Affecting the Purchase Decision of Investors in the Residential Property Market in Malaysia," *J. Surv. Constr. Prop.*, vol. 5, no. 2, pp. 1–13, 2014.

[3] N. Vineeth, M. Ayyappa, and B. Bharathi, "House Price Prediction Using Various Machine Learning Algorithms," in *Communications in Computer and Information Science*, 2018, vol. 837, pp. 425–433.

[4] A. Nguyen, C. Fernandes, N. Webb, and H. Holt, "Housing Price Prediction," Union College, 2018.

[5] S. Ng, "Steps to buying a home," *EdgeProp.my*, 2019. [Online]. Available: https://www.edgeprop.my/content/1485022/steps-buying-home. [Accessed: 01-Jul-2020].

[6] A. S. Ravikumar, "Real Estate Price Prediction Using Machine Learning," National College of Ireland, 2018.

[7] R. Bafna, A. Dhole, A. Jagtap, A. Kazi, and A. Kazi, "Prediction of Residential Property Prices – A State of the Art," *Int. Adv. Res. J. Sci. Eng. Technol.*, vol. 5, no. May, pp. 2007–2010, 2018.

[8] T. Mohd, S. Masrom, and N. Johari, "Machine learning housing price prediction in Petaling Jaya, Selangor, Malaysia," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special Issue 11, pp. 542–546, 2019.

[9] S.-H. Chunga, H.-L. Mab, M. Hansend, and T.-M. Choi, "Transportation Research Part E," *Transp. Res. Part E*, vol. 134, no. January, 2020.

[10] N. Pow, E. Janulewicz, and L. (Dave) Liu, "Applied Machine Learning Project 4 Prediction of real estate property prices in Montreal," pp. 1–8, 2014.

[11] J. Hong, H. Choi, and W. Kim, "A house price valuation based on the random forest approach : The mass

appraisal of residential property in South Korea," *Int. J. Strateg. Prop. Manag.*, pp. 1–13, 2020.

[12] M. Thamarai and S. P. Malarvizhi, "House Price Prediction Using Machine Learning," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9, pp. 717–722, 2019.

[13] T. D. Phan, "Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia," *Proc. - Int. Conf. Mach. Learn. Data Eng. iCMLDE 2018*, pp. 8–13, 2019.

[14] S. N. A. Rahman, N. H. A. Maimun, M. N. Razali, and S. Ismail, "The artificial neural network model (ANN) for Malaysian housing market analysis," *Plan. Malaysia*, vol. 17, no. 1, pp. 1–9, 2019.

[15] T. Fu, "Forecasting Second-hand Housing Price using Artificial Intelligence and Machine Learning Techniques," *8th Int. Conf. Mechatronics, Comput. Educ. Informationization (MCEI 2018)*, vol. 83, no. Mcei, 2018.

[16] M. F. Mukhlishin, R. Saputra, and A. Wibowo, "Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor," *Proc. - 2017 1st Int. Conf. Informatics Comput. Sci. ICICoS 2017*, no. 1, pp. 171–176, 2018.

[17] T. Dimopoulos, H. Tyralis, N. P. Bakas, and D. Hadjimitsis, "Accuracy measurement of Random Forests and Linear Regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus," *Adv. Geosci.*, vol. 45, pp. 377–382, 2018.

[18] Y. Ma, Z. Zhang, A. Ihler, and B. Pan, "Estimating warehouse rental price using machine learning techniques," *Int. J. Comput. Commun. Control*, vol. 13, no. 2, pp. 235–250, 2018.

[19] S. Borde, A. Rane, G. Shende, and S. Shetty, "Real Estate Investment Advising Using Machine Learning," *Int. Res. J. Eng. Technol.*, vol. 4, no. 3, pp. 1821–1825, 2017.

[20] J. Y. Wu, "Housing Price prediction Using Support Vector Regression," San Jose State University, 2017.

[21] M. A. Valle, R. Crespo, A. A. V Schuler, and F. Crespo, "Property Valuation using Machine Learning Algorithms," *Int. Conf. Model. Simul.*, no. January, pp. 97–105, 2016.

[22] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, "A hybrid regression technique for house prices prediction," *IEEE Int. Conf. Ind. Eng. Eng. Manag.*, vol. 2017-Decem, pp. 319–323, 2018.

[23] A. Komagome-towne, "Models and Visualizations for Housing Price Prediction," Faculty of California State Polytechnic University, Pomona, 2016.

[24] "Pricing brickz," *Brickz Research Sdn Bhd*, 2020. [Online]. Available: https://www.brickz.my/pricing/. [Accessed: 25-Jun-2020].

[25] H. Brink, J. W. Richards, and M. Fetherolf, *Real-World Machine Learning*. Manning Publication Co., 2017.

[26] S. Abbasi, "Advanced Regression Techniques Based Housing Price Prediction Model," *13th Int. Conf. Iran. Oper. Res. Soc.*, no. February, pp. 1–10, 2020.

[27] "Penang - Wikipedia," *Wikipedia*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Penang#Local_governments. [Accessed: 26-Jun-2020].

[28] R. K. Paul, "Multicollinearity: causes, effects and remedies," *Indian Agric. Stat. Res. Inst.*, no. 4405, 2014.

[29] J. I. Daoud, "Multicollinearity and Regression Analysis," *J. Phys. Conf. Ser.*, vol. 949, no. 1, 2018.

[30] S. Raschka and V. Mirjalili, *Python Machine Learning - Machine Learning and Deep Learning with Python, scikit-learn and TensorFlow*, Second Edi. Packt Publishing, 2017.

[31] G. Shmueli, P. C. Bruce, P. Gedeck, and N. R. Patel, *Data Mining for Business Analytics. Concepts, Techniques and Applications in Python*, 1st ed. John Wiley & Sons, Inc., 2020.

[32] "Linear Regression Tool," *Alteryx Documentation*, 2020. [Online]. Available: https://help.alteryx.com/current/designer/linear-regression-tool. [Accessed: 27-Nov-2020].

[33] "Forest Model Tool," *Alteryx Documentation*, 2020. [Online]. Available: https://help.alteryx.com/current/designer/forest-model-tool. [Accessed: 27-Nov-2020].

[34] H. Yu and J. Wu, "Real Estate Price Prediction with Regression and Classification," *CS 229 Autumn 2016 Proj. Final Rep.*, pp. 1–5, 2016.

[35] "Find Nearest Neighbors Tool," *Alteryx Documentation*, 2020. [Online]. Available: https://help.alteryx.com/current/designer/find-nearest-neighbors-tool. [Accessed: 27-Nov-2020].

[36] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*, vol. 9780470908. 2014.

[37] "Cover Tree," *Wikipedia*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Cover_tree. [Accessed: 03-Dec-2020].

[38] "k-d Tree," *Wikipedia*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/K-d_tree. [Accessed: 03-Dec-2020].

[39] W. N. Venables and B. D. Ripley, "Tree-Based Methods," in *Modern Applied Statistics with S, 4th ed.*, Springer, Berlin, 2002, pp. 251–269.

[40] S. Tarang, "About Train, Validation and Test Sets in Machine Learning," 2017. [Online]. Available: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7. [Accessed: 31-Mar-2020].

[41] "Northeast Penang Island District - Wikipedia," *Wikimedia Foundation, Inc.*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Northeast_Penang_Island_District. [Accessed: 11-Nov-2020].